

SPEECH TRANSCRIPTION TOOL
FOR EFFICIENT SPEECH TRANSCRIPTION

RELATED APPLICATION

[0001] This application is related to the concurrently-filed U.S. application (Docket No. 02-4040), serial number _____, titled "Fast Transcription of Speech," which is incorporated herein by reference.

[0002] This application claims priority under 35 U.S.C. § 119 based on U.S. Provisional Application No 60/419,214 filed October 17, 2002, the disclosure of which is incorporated herein by reference.

GOVERNMENT CONTRACT

[0003] The U.S. Government has a paid-up license in this invention and the right in limited circumstances to require the patent owner to license others on reason-able terms as provided for by the terms of (contract No. 1999*S018900*000) awarded by Federal Broadcast Information Service (FBIS).

BACKGROUND OF THE INVENTION

A. Field of the Invention

[0004] The present invention relates generally to speech processing and, more particularly, to the transcription of speech.

B. Description of Related Art

[0005] Speech has not traditionally been valued as an archival information source. As effective as the spoken word is for communicating, archiving spoken segments in a useful and easily retrievable manner has long been a difficult proposition. Although the act of recording audio is not difficult, automatically transcribing and indexing speech in an intelligent and useful manner can be difficult.

[0006] Automatic transcription systems are generally based on a language model. The language model is trained on a speech signal and on a corresponding transcription of the speech. The model will “learn” how the speech signal corresponds to the transcription. Typically, the training transcriptions of the speech are derived through a manual transcription process in which a user listens to the training audio and types in the text corresponding to the audio.

[0007] Manually transcribing speech can be a time consuming and, thus, expensive task. Conventionally, generating one hour of transcribed training data requires up to 40 hours of a skilled transcriber’s time. Accordingly, in situations in which a lot of training data is required, or in which a number of different languages are to be modeled, the cost of obtaining the training data can be prohibitive.

[0008] Thus, there is a need in the art to be able to cost-effectively transcribe speech.

SUMMARY OF THE INVENTION

[0009] Systems and methods consistent with the principles of this invention provide a transcription tool that allows a user to efficiently transcribe segments of speech to generate a structured and annotated transcription.

[0010] One aspect consistent with the invention is directed to a speech transcription tool. The speech transcription tool includes control logic, an input device, and a graphical user interface. The control logic is configured to play back portions of an audio stream and the input device receives text from a user defining a transcription of the portions of the audio stream and receives annotation information from the user further defining the text. The graphical user interface includes a first section that displays a graphical representation of a waveform corresponding to the audio stream and a second section that displays the text and representations of the annotation information for the text.

[0011] A second aspect consistent with the invention is directed to a method that comprises receiving an audio stream containing speech data, receiving text from a user defining a transcription of the speech data, receiving annotation information from the user further defining the text, displaying the text, and displaying symbolic representations of the annotation information with the text.

[0012] A third aspect consistent with the invention is directed to a computing device for transcribing an audio file that includes speech. The computing device includes an audio output device, a processor, and a computer memory. The computer memory is coupled to the processor and contains programming instructions that when executed by the processor cause the processor to play a

current segment of the audio file through the audio output device, receive transcription information for the speech segments played through the audio output device, receive annotation information relating to the transcription information, and display the transcription information in an output section of a graphical user interface. Additionally, the processor displays the annotation information as graphical icons in the output section of the graphical user interface.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the invention and, together with the description, explain the invention. In the drawings,

[0014] Fig. 1 is a diagram illustrating an exemplary system in which concepts consistent with the invention may be implemented;

[0015] Fig. 2 is a block diagram of a transcription tool consistent with the present invention;

[0016] Fig. 3 is an exemplary diagram of an interface that may be presented to the user of the transcription tool shown in Fig. 2;

[0017] Fig. 4 is a flow chart illustrating exemplary operation of the transcription tool shown in Fig. 2;

[0018] Fig. 5 is a diagram illustrating user selection of a speaker turn; and

[0019] Fig. 6 is an exemplary diagram of an interface including a pop-up box for further defining annotations.

DETAILED DESCRIPTION

[0020] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers may be used in different drawings to identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents of the claim limitations.

[0021] A speech transcription tool assists a user in transcribing speech. The speech transcription tool allows the user to transcribe and annotate speech in intuitive ways. The transcription tool presents an integrated view to the user that includes a view of the audio waveform, a view of the text input by the user, and a view of the structured version of the transcribed text. The view of the text input by the user may include graphical icons that represent annotation information that relates to the transcribed text.

SYSTEM OVERVIEW

[0022] Speech transcription, as described herein, may be performed on one or more processing devices or networks of processing devices. Fig. 1 is a diagram illustrating an exemplary system 100 in which concepts consistent with the invention may be implemented. System 100 includes a computing device 101 that has a computer-readable medium, such as a random access memory 109, coupled to a processor 108. Computing device 101 may also include a number of additional external or internal devices. An external input device 120 and an external output device 121 are shown in Fig. 1. The input device 120

may include, without limitation, a mouse, a CD-ROM, or a keyboard. The output device may include, without limitation, a display or an audio output device, such as a speaker. A keyboard, in particular, may be used by the user of system 100 when transcribing a speech segment that is played back from an output device, such as a speaker. A foot pedal may be used for audio playback control.

[0023] In general, computing device 101 may be any type of computing platform, and may be connected to a network 102. Computing device 101 is exemplary only. Concepts consistent with the present invention can be implemented on any computing device, whether or not connected to a network.

[0024] Processor 108 executes program instructions stored in memory 109. Processor 108 can include any of a number of well-known computer processors, such as processors from Intel Corporation, of Santa Clara, California.

[0025] Memory 109 contains an application program. In particular, the application program may implement a transcription tool 115 described below. Transcription tool 115 plays audio segments to a user. The user transcribes speech in the audio and enters annotation information that further describes the transcription into transcription tool 115.

TRANSCRIPTION TOOL

[0026] Fig. 2 is a block diagram illustrating software elements of transcription tool 115. Users of transcription tool 115 (i.e., transcribers) interact with transcription tool 115 through user input component 203 and graphical user interface (GUI) 204. Control logic 202 coordinates the operation of graphical

user interface 204 and user input component 203 to perform transcription in a manner consistent with the present invention. Control logic 202 may additionally handle the playback of the input audio to the user.

[0027] User input component 203 processes information received from the user. A user may input information through a number of different hardware input devices. A keyboard, for example, is an input device that the user may use in entering text corresponding to speech. Other devices, such as a foot pedal or a mouse, may be used to control the operation of transcription tool 115.

[0028] Graphical user interface 204 displays the graphical interface through which the user interacts with transcription tool 115. Fig. 3 is an exemplary diagram of an interface 300 that may be presented to the user via graphical user interface 204. Interface 300 includes waveform section 301, transcription section 302, and structured representation section 303. Interface 300 may include selectable menu options 304 and window control buttons 305. Through menu options 304, a user may initiate functions of transcription tool 115, such as opening an audio file for transcription, saving a transcription, and setting program options.

[0029] Waveform section 301 graphically illustrates the time-domain waveform of the audio stream that is being processed. The exemplary waveform shown in Fig. 3, waveform 310, includes a number of quiet segments 311 and audible segments 312. Audible segments 312 may include, for example, speech, music, other sounds, or combinations thereof.

[0030] Concurrently with the display of audio waveform 310, transcription tool 115 may play the audio signal to the user. Transcription tool 115 may visually mark the portion of waveform 310 that is currently being played. For example, as shown in Fig. 3, an arrow 316 may point to the current playback position in audio waveform 310. The user may move the arrow using a mouse or keyboard commands to quickly adjust the current playback position.

[0031] Sections of waveform 310 may be labeled as corresponding to different segments of an audio stream. The segments may be defined hierarchically. In one implementation consistent with the invention, these different segments may include “turns,” “sections,” and “episodes.” Additional segments, such as a “gap” segment that defines a period of non-speech such as silence, music, noise, etc., may also be used. A turn may refer to a section of the audio in which a single speaker is speaking (i.e., a “speaker turn”). A section may refer to a number of speaker turns that relate to a particular topic. An episode may refer to a group of sections that each have something in common, such as being of the same broadcast. The turn, section, and episode segments for an audio stream may be illustrated in waveform section 310 using graphical markers. In Fig. 3, these three segments, as well as the gap segment, are illustrated as turns 320, sections 321, episodes 322, and gaps 323.

[0032] One type of audio stream that can be confidently divided into turns, sections, and episodes is a news broadcast. The whole news broadcast (e.g., a 30 minute broadcast) may correspond to a single episode. An episode may

include multiple sections that each corresponds to a different news story. Each section may have one or more speaker turns.

[0033] Transcription section 302 displays transcribed text received by control logic 202 from user input component 203. The text may be represented in Unicode so that the transcription tool can handle left-to-right, right-to-left, and bi-directional scripts, such as English, Chinese, and Arabic. Typically, the text will be typed by the user as the user listens to the audio waveform 310. In addition to merely typing the text of the transcription, the user may input additional information relating to the text. This additional information is received by control logic 202 and stored as annotation information for the text. Annotations may include, for example, an indication that a certain noun corresponds to a person's name or to a location. The annotation information may be displayed in transcription section 302. Annotations 313 and 314 are shown in Fig. 3 that each defines a word or series of words. More particularly, annotations 313 define names of persons and annotations 314 define location names. Annotations may additionally be nested. For example, in the phrase "CNN News," "CNN" may be annotated as "spelled" and the complete phrase "CNN News" may be a "name" annotation.

[0034] Structured representation section 303 displays the transcribed text in a hierarchical tree structure. The hierarchical structure may be based on the relationships of segments 320-322. Thus, in Fig. 3, for example, an episode entry 330 (e.g., a folder icon) is at the highest level. An episode may include one or more section entries 331, which may include one or more turn and/or gap

entries 332. The turn entries are at the base level in the hierarchy and contain the actual transcription text. One of turn entries 332 is highlighted in Fig. 3, which indicates that this is the currently active turn. Turn entry 332, in addition to the transcribed text, may include annotation information that was input by the user, such as the name of the speaker ("Riaz Ahmad Khan") and the sex of the speaker ("male"). The sex of the speaker may alternatively be determined automatically based on acoustic processing techniques applied to the speaker turn. Section entries 331 may include a general description of the topic(s) discussed in the turns corresponding to a section. The topic description may be determined automatically based on the speaker turn transcriptions.

[0035] When the user finishes a transcription, transcription tool 115 may save the transcription as an output file. The output file may be based on the information in structured representation section 303. That is, the output file may include the transcribed text as well as meta-data that encapsulates the annotation information, including indications of the hierarchical segments. In one implementation, the output file may be an extensible markup language (XML) document.

[0036] Fig. 4 is a flow chart illustrating exemplary operation of transcription tool 115 consistent with an aspect of the invention. Before transcribing speech, the user may first load an audio waveform into transcription tool 115. This may be accomplished through the "file" menu.

[0037] With the waveform loaded, the user may define segments in the waveform, such as turn, section, and episode segments (Act 401). Segments

may be defined by, for example, using a mouse to highlight a continuous portion of the audio that corresponds to a single speaker. Fig. 5 is a diagram illustrating a waveform in which the user has highlighted a portion 501 (shown as a simple rectangle in Fig. 5) that corresponds to a speaker turn. The highlighted portion 501 may include buffer areas 502 that the user aligns to the edge of the speaker turn. Transcription tool may 115 adjust the graphical marker 520 that defines the speaker turn as the user varies the highlighted portion with the mouse. When the user has adjusted highlighted portion 501 to adequately cover the speaker turn, the user may press a predefined key combination, such as CTRL-T, that causes control logic 202 to store the speaker turn. Other user actions, such as a mouse click, instead of a keyboard combination, may be used to inform control logic 202 of a speaker turn.

[0038] In some implementations, the user may load a saved version of the waveform in which segments have already been defined. In this situation, the user may not have to re-define the segments.

[0039] The user may define sections and episodes in a manner similar to defining speaker turns. Alternatively, control logic 202 may automatically define sections and/or episodes based on the transcribed context of the speaker turns. Control logic 202 may, for example, determine that speaker turns are similar based on the text of the speaker turns. Speaker turns discussing the same topic will tend to use similar words and may, thus, be compared for similarity based on the frequency of occurrence of words in the speaker turn.

[0040] Additionally, instead of having the user manually highlight portions 501 of a speaker turn, control logic 202 may use automated speech and language processing functions to initially classify the audio based on an audio type, such as speech, music, or silence. One such technique for automatically identifying segments in an audio stream, such as speaker turns, appropriate for transcription are discussed in the application cited in the “Related Application” section of this document. Portions of the audio that contain only music may be noted on interface 300 so that the user does not need to bother listening to these portions.

[0041] As the user defines speaker turns, sections, and episodes, transcription tool 115 may update structured representation section 303 to indicate the defined segments. The user may listen to audio before creating segments to determine where turn boundaries should be.

[0042] After defining one or a number of segments, the user may begin playback of a particular one of the segments, such as a speaker turn (Act 402). In one implementation, the user may control which of the speaker turns is the active speaker turn via mouse or keyboard commands. Thus, a user may point to a particular speaker turn 320 to select the corresponding section of waveform 310 for playback. Alternatively, the user may adjust the current playback position using predefined keyboard commands. For example, the key combination CTRL-↓ may cause control logic 202 to select the next speaker turn as the active speaker turn, the key combination CTRL-↑ may cause control logic 202 to select the previous speaker turn as the active speaker turn, and the key combinations

SHIFT-CTRL-←/→ may move the current active location, as indicated by arrow 316, to the left/right in predetermined increments (e.g., 0.8 second increments).

[0043] While transcription tool 115 is playing back audio, the user may transcribe speech in the audio by entering (e.g., typing) the text into user input device 203 (Act 403). Control logic 202 displays the text in transcription section 302 and may simultaneously update structured representation section 303. During the transcription process, the user may enter annotation information for a particular word or sequence of words (Acts 404 and 405).

[0044] In one implementation, the annotation information is entered by a user through keyboard shortcuts. For example, before typing in a name, the user may input a key combination such as CTRL-N. This key combination informs control logic 202 that the succeeding text corresponds to a name. In some implementations, pressing CTRL-N may bring up a selection box that allows the user to further define the name that is to be annotated, such as the name of a person or the name of a location. Fig. 6 is an exemplary diagram of an interface 600 for transcription tool 115 that includes a pop-up box 601 that allows the user to further define name annotations. Control logic 202 may display pop-up box 601 in response to the keyboard combination (e.g., CTRL-N) for a name object.

[0045] Based on the selected name, control logic 202 may generate an appropriate name icon surrounding text typed by the user, such as name icons 313 or 314 (Fig. 3). When the user has completed typing the name, he may again press CTRL-N to turn off name annotation and revert back to normal text transcription.

[0046] Other annotations, in addition to name annotations, may be entered by the user. For example, the user may mark an unintelligible section of speech with a “skip” marker that is toggled on/off via the key combination CTRL-K.

[0047] When the user finishes transcribing (Act 406), transcription tool 115 may output the transcription entered by the user to a file (Act 407). In one implementation, the user selects the output file to write to using the “File” menu on interface 300. As previously mentioned, the output file may be an XML document that includes the information in structured representation section 303. Thus, the output file may include the transcribed text, the annotation information, the segmentation information, and other information, such as time codes that correlate the transcription with the original audio.

[0048] The output file, in addition to being an XML document, may be generated using Unicode to represent the characters. The Unicode standard is a character encoding standard that represents characters using a unique number for every character, regardless of the computing platform or language. The Unicode standard is maintained by the Unicode consortium.

[0049] In addition to entering information while transcribing text, in some implementations, users may enter annotation information after transcribing the text. In particular, control logic 202 may allow users to highlight text in transcription section 302 and then select the annotation information to apply to the highlighted text.

TRANSCRIPTION TOOL CONFIGURATION

[0050] When initially starting up, transcription tool 115 may read a configuration file. The configuration file may define functionality for a number of operational aspects of the transcription tool. For example, the configuration file may define the names and the relationships (e.g., hierarchy) between the segments, the possible annotation information, and the keyboard shortcuts that are used to enter the annotation information. In this manner, by modifying the configuration file, transcription tool 115 can be customized for a particular transcription task.

[0051] In one implementation, the structure of the configuration file is defined through an XML schema definition.

CONCLUSION

[0052] The transcription tool described herein allows users to efficiently create rich transcriptions of an audio stream. In addition to merely typing in spoken words, the user may easily annotate the spoken words and segment the audio stream into useful segments. Moreover, the categories of allowed annotation information and the possible segments can be easily modified by the user by changing a configuration file.

[0053] The foregoing description of preferred embodiments of the invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are

possible in light of the above teachings or may be acquired from practice of the invention. For example, while a series of acts has been presented with respect to Fig. 4, the order of the acts may be different in other implementations consistent with the present invention. Also, certain actions have been described as keyboard actions, however, these actions might also be performed via other input devices such as a mouse or a footpedal.

[0054] Certain portions of the invention have been described as software that performs one or more functions. The software may more generally be implemented as any type of logic. This logic may include hardware, such as an application specific integrated circuit or a field programmable gate array, software, or a combination of hardware and software.

[0055] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article “a” is intended to include one or more items. Where only one item is intended, the term “one” or similar language is used.

[0056] The scope of the invention is defined by the claims and their equivalents.